基于突显词博文聚类的官微事件检测方法*

高永兵[1]　杨贵朋[1]　张　娣[1]　马占飞[2]
[1](内蒙古科技大学信息工程学院　包头 014010)
[2](包头师范学院计算机系　包头 014010)

摘要: 【目的】针对官方微博数据存在大量不相关信息的问题，过滤博文进而检测事件。【方法】利用 Word2Vec 机器学习模型训练官方微博记录集，并将博文影响力、词基础权重以及官微相关性相结合，提出官方微博突显词检测方法，计算突显词博文的相似度，利用层次聚类算法对突显词博文聚类后选取合适的突显词描述事件，从而实现事件检测。【结果】实验结果表明，与 TF-IDF 和 TextRank 算法相比较，本文的突显词算法在准确率(63.5%)、召回率(85.5%)和 F 值(73.0%)方面表现更好。【局限】官方微博历史记录太少，初始的训练会存在数据冷启动问题。【结论】本文方法可以在官方微博博文中有效检测官方微博事件。

关键词: 官方微博　相关词　突显词　官微事件　Word2Vec
分类号: TP391　G35

1　引　言

随着大量组织机构微博的开通，官方微博(简称官微)逐渐进入人们的视野，并引起社会各界的重视，也引发了学术界的研究热情。官微一般属于组织团体账号，是经过平台认证的微博。官微博文比较正式，可信度高，其组织功能的宣传性博文占比大，具有较强的社会效应。但由于官微多人式分工维护的特点，博文中蕴含着大量官微发展历程信息，也存在许多非组织功能性的博文，以新浪微博北京大学官微中两条博文为例:

P1: #好读书，读好书#【每周读书特辑•外国小说】孟德斯鸠说: 喜爱...外国小说更是给我们呈现出一个完全不同的世界。

P2: #总理来啦# 第一站，克强总理来到位于朗润园的国家发展研究院，了解北京大学智库建设以及国家发展研究院的发展情况。北京大学校长林建华、校党委书记朱善璐陪同视察...。

笔者定义官微事件为描述官方组织机构在某段时间内所发生的事情，或者与其相关的事情。P2 谈及总理来北京大学访问的事件，就是比较受关注的北京大学官微事件。本文目标是过滤掉类似 P1 的博文，保留类似 P2 的博文并从过滤后的博文中聚类时间上相近、内容上相关的博文，进而实现官微事件检测，并筛选出特征词描述所检测到的官微事件。在官微逐渐成为团体组织和企业主要宣传阵地的今天，从官微历史记录中提取官微事件有助于浏览者快速了解该组织的主要事件，大大提高信息获取效率，但如何过滤官微博文中的无关组织功能性博文、准确提取事件特征词的工作存在着不少挑战。

2　相关工作

微博事件检测技术一直以来都是学术界备受关注的研究热点，其基本思路是通过检测具有热点时间突发效应的高频词汇并计算语句间的相似性，将相关度高的大量语句段落聚合到一起，通过句法分析和词性分析提取事件[1]。

目前的研究工作主要是通过微博话题或主题的检测分析进行事件检测[2-3]，文献[4]提出一种结合微博

通讯作者: 高永兵, ORCID: 0000-0002-9950-7391, E-mail: gaoyongbing@126.com。

*本文系国家自然科学基金项目“面向物联网安全的 Multi-ISM 协同建模及关键技术研究”(项目编号: 61163025)和内蒙古自然科学基金项目“基于个人微博的自动摘要关键技术研究”(项目编号: 2015MS0621)的研究成果之一。

数据文本特征、语义特征、时序特征和社交关系特性的微博数据事件检测算法 EDM，与 LDA 模型的事件检测算法对比显示出其事件检测算法的有效性。文献[5]提出基于突发词特征增量聚类的微博新闻话题检测方法。该方法引入突发词特征增量聚类算法对新闻话题进行发掘，其具有较高的算法效率，但忽略了微博发文率等博文特有特征，而且只考虑词频增长率等微博文本信息抽取事件突发特征。事件检测方法[6-10]通过构造词汇文本矩阵分析事件，微博数据的短文本和文本缺失性导致特征矩阵高度稀疏，实验结果的准确率难以令人满意。另外，微博数据的转发评论为事件检测提取提供丰富的数据基础，传统的方法未将其考虑进去。目前的研究成果大多基于改进的 TF-IDF、基于概率和基于图的方法。近年来，机器学习、大数据处理等领域方法在微博数据研究方面崭露头角，譬如基于 Hadoop 框架分布式处理微博数据、分布式词向量概念的推出等都对微博研究起到很大的推动作用。以往工作主要针对于公共微博集的公共事件做深入研究，就单个官微事件检测的研究还比较欠缺，官微事件的研究工作仍需要进一步完善。本文针对官微事件检测提出如下定义。

定义1 相关词：与某一检索词有如下关系的其他检索词。

①所属关系。如："北京大学"、"博雅塔""农园食堂"。

②关联关系。如："北京大学"、"空间科学""清华大学"。

官微相关词是能够反映官微组织事件主题信息的相关词，这些词通常在官微博文历史记录里拥有较高的词频，同时与官微事件存在一定的联系。

定义2 突显词：在某个或多个不同时间段内，该词经常出现，且上述时间段以外很少出现或不出现的特征实词。

①屠呦呦获奖时间段。如："屠呦呦"、"诺奖"、"林建华"。

②总理来访时间段。如："总理"、"林建华"、"农园食堂"。

官微突显词是博文中能够描述官微事件的具有一定影响力、相关性的突显词。

本文综合考虑官微博文特征，借助机器学习模型训练官微相关词，并基于官微相关词、时间段博文活跃度以及博文的转发评论等基础特征提取官微突显词，计算官微突显词博文相似度并聚类检测官微事件。选取能描述官微事件的突显词实现官微事件检测。

3 官微突显词提取

3.1 相关词训练

根据官方微博的数据流特征引入官方微博相关语料集，通过官微相关词权重加权可以提高官微博文突显词提取的准确性。借助程序抓取官微组织的微博历史记录集，利用 Word2Vec[11]工具训练并建立官微相关语料库。

Word2Vec 是一款将词表征为实数值向量的高效工具，其利用深度学习思想进行记录集的训练。Word2Vec 使用的是分布式词向量表示方式，基本思想是通过训练将每个词映射成 N 维实数向量(N 一般为模型中的超参数)，通过词之间的余弦相似度距离、欧式距离等距离计算方法判断词之间的语义相似度。其核心架构由 CBOW 和 Skip-gram 两个模型组成，采用输入层、隐藏层、输出层的三层神经网络对语言模型进行建模，同时获得一种单词在向量空间上的表示。核心技术是根据词频 Huffman 编码，使得所有词频相似的词隐藏层激活的内容基本一致，出现频率越高的词或者词语，激活的隐藏层数目越少，这样有效地降低了计算的复杂度。Word2Vec 输出的词向量可以被用来进行很多自然语言处理方面的相关工作，譬如聚类、同义词识别、词性分析等。与潜在语义分析 LSI、潜在狄利克雷分布 LDA 的经典过程相比，Word2Vec 利用词的上下文信息，语义信息更加的全面、丰富。本文利用 Word2Vec 的词向量模块训练记录集，选取 Skip-gram 模型输入官微名称，得到官微相关词和相关度权值并构成官微相关语料集。语料集中的元素由相关词 $word$ 和相关度权值 w_ra 组成，如公式(1)所示。

$$REL_{i,j}=\{(word_1,w\_ra_1),(word_2,w\_ra_2),\cdots (word_m,w\_ra_m)\cdots\} \quad (1)$$

其中，$REL_{i,j}$ 表示官微相关语料集，$word_m$ 和 w_ra_m 分别表示官微相关词和相关度权值。

3.2 突显词提取

通过检测官微博文中特征词的突显性提取官微突显词。博文中特征词的突显性与博文影响力[12-13]、词基础权重和官微相关度权值有关。

(1) 博文影响力

官微特征词的突显性与官微博文影响力之间存在很大关联。描述官微事件的突显词是具备影响力的，

这些词在自己的微博博文中不仅经常被提及，而且在其他微博用户的博文中也能够见到。不同于传统网页文本，譬如长文本文档，一个词汇会引起更多人的注意可能是其出现在标题或者版面上，也意味着其影响力很大。但是微博受到字符长度限制，也不存在标题和版面，只有用户关注的博文才会被浏览。因此考虑到官微博文的转发评论、官微博文活跃度等特性。笔者将时间段 t 中包含特征词 i 的官微博文 P 的影响力表示为如公式(2)所示。

$$INF_i^t(P)=\sum_{i\in P}^{t}\lg(nct_P+nrt_P+1)\times V_P \qquad (2)$$

其中，P 表示包含特征词 i 的所有博文，nct_P 和 nrt_P 分别表示包含特征词 i 的博文的被转发数和被评论数。V_P 表示官微的博文活跃度。官微博文的活跃度由官微博文某一段时间内的发博频率，即一段时间内平均每天发布博文的条数决定，该频率能够反映出官微某段时间内的博文活跃程度，如公式(3)所示。

$$V_P=\begin{cases}0.5 & avgP\geqslant 15\\ 0.3 & 3\leqslant avgP<15\\ 0 & avgP<3\end{cases} \qquad (3)$$

其中，V_P 表示官微博文活跃度，$avgP$ 表示官微组织一段时间 t 内平均每天的发博率，即官微时间 t 内平均每天发布的微博条数，该值越大，意味着官微博文在此段时间内越活跃，则官微事件出现的可能性就越大。

(2) 词基础权重

在突显词检测时，借助于 TextRank[14-15]算法获取博文中的关键词权重作为特征词的基础权重值 $TW(V_i)$。TextRank 是基于图的原理，针对文本句子关系设计的权重算法，将词语视为句子，通过构建词语图将每个词作为节点，边作为权重，基于相邻词进行投票的原理，利用局部词汇之间关系对后续特征词进行排序，直接从文本本身抽取。因其简洁有效而得到广泛使用。计算方法如公式(4)所示。

$$TW(V_i)=(1-d)+d\times\sum_{V_j\in In(V_i)}\frac{w_{ji}}{\sum_{V_k\in Out(V_j)}w_{jk}}TW(V_j) \qquad (4)$$

其中，$TW(V_i)$表示特征词 i 的权重，d 为阻尼系数，一般取值 0.85; $In(V_i)$表示节点 V_i 的邻接边集合即指向词 i 的词语集合，$Out(V_j)$表示节点 V_j 的相邻边集合即指向词 j 的词语集合，w_{ji} 和 w_{jk} 表示边权重。

(3) 官微突显词

官微突显词是具有突显性的官微特征词，对发现新兴官微事件的研究工作尤为重要。官微特征词的突显性在日常中表征较为普遍，可以和任何官微事件保持独立，但当某一重要官微事件发生时，该词的影响力、突显性是呈现增大状态的。考虑到官微的复杂结构，综合博文影响力、基础权重和官微相关词权重等因素来评价官微特征词的突显性，如公式(5)所示。

$$Burst_i^t=\frac{1}{n}\sum_{k=t-n}^{t-1}[(INF_i^t(P)-INF_i^k(P))\times TW(V_i)\times(1+REL_{i,j})] \qquad (5)$$

其中，$Burst_i^t$ 是官微特征词 i 在时间段 t 内的官微突显度，n 是时间段 t 之前的 n 个时间段的大小，取值是在 1 到 t 之间的整数。$INF_i^k(P)$表示时间段 k 内包含特征词 i 的博文的影响力大小; $REL_{i,j}$ 表示词 i 与官微的相关度权值，该值通过 Word2Vec 中模型训练所得，未训练出的官微相关词相关度权值为 0。根据官微特征词突显性筛选规则，抽取时间段 t 内所有突显性大于阈值 σ 的官微特征词，称之为官微突显词，突显词构成的集合为官微突显词集，官微突显词集的表示如公式(6)所示。

$$Burst^t=\{theme_1^t,\ theme_2^t,\cdots,theme_i^t,\cdots\} \qquad (6)$$

其中，$theme_i^t$ 表示时间窗 t 内的第 i 个官微突显词。

官微突显词具有较高的官微事件主题表现力，能够体现官微事件的突显性。即其与官微事件是紧密相关的，借助它可以概括整个官微事件，也可以描述官微事件的某一个方面，表征官微事件。

4　基于突显词博文聚类的官微事件检测

对官微特征词计算突显性后，得到官微突显词。也就是官微事件的相关官微博文被表示成无权重的官微突显词特征集合。通过计算含有突显词博文的综合相似度用于官微突显词博文的聚类，利用相应的官微突显词簇表征官微事件类。相似度计算方法采用基于语义特征的微博相似度计算方法[16]，其中文本语义同时考虑知网语义(包括义原、义项相似度)，并结合词频的因素[17]，能够将相关度高的突显词博文归为一类，又避免不相关博文带来的影响。文本语义集成标签、链接标题和时间相似性计算综合相似度如公式(7)–公

式(9)所示。

$$Simd_{pt} = Sim_{Semantic}(p_i, p_k) + Sim_{tf\text{-}idf}(p_i, p_k) \quad (7)$$

$$Simd_{ts} = Sim_{ts}(p_i, p_k) = \exp(-\frac{|t_i - t_k|}{\theta}) \quad (8)$$

$$Simd(p_i, p_k) = \alpha_1 Simd_{pt} + \alpha_2 Simd_{url} + \alpha_3 Simd_{tag} + \alpha_4 Simd_{ts} \quad (9)$$

其中，$Sim_{Semantic}$ 是含突显词博文的语义相关度，$Sim_{tf\text{-}idf}$ 是其 tf-idf 相关度；$Simd(p_i, p_k)$ 表示含官微突显词的博文 p_i 和 p_k 之间的综合相似度，$Simd_{pt}$、$Simd_{url}$、$Simd_{tag}$ 和 $Simd_{ts}$ 分别表示含有官微突显词的博文纯文本相似度、链接标题相似度、标签相似度和时间相似度；对于链接标题和标签相似度，如果词语被知网收录，采用语义相关度计算，否则直接利用词频向量进行计算；由实验可知，时间相似度中参数 θ 取 0.3 时精确度最高。其中 α_1、α_2、α_3、α_4 参数值分别设置为 0.3、0.3、0.25、0.15 时，人工标注下的相似官微博文的相似度测试结果最好。

通过计算官微中任意两个官微突显词博文之间相似性得到一个无向图 $G(V, E)$，顶点 $v_i \in V$ 表示突显词博文 p_i、p_i 和 p_k 之间的边 $e_{i,k} \in E$ 的权重值表示 p_i 和 p_k 之间的综合相似性。在此无向图上采用凝聚层次聚类算法 HAC 对图进行聚类，将得到的类作为官微事件类的检测结果。通过类中的官微突显词描述官微事件，进而生成官微事件检测结果。聚类及事件检测算法过程如下：

```
输入：含突显词官微博文集合 P_Burst';
输出：含突显词官微博文类簇集 Clusters。
    simMatrix={}
    for P_Burst' 的每个突显词博文 p_i do
        for P_Burst' 的每个突显词博文 p_k do
            simMatrix=[i][k]=simd(p_i, p_k)
        end for
    end for
    Clusters =clusterbyHAC(simMatrix, P_Burst')
    return Clusters
```

算法定义相似度矩阵，计算出任意两个突显词博文簇之间的相似性，利用凝聚层次聚类算法 HAC 进行聚类，得到官微博文事件类簇结果。算法的平均时间复杂度为 $O(n2)$。HAC 算法是一种自底向上的凝聚式算法，其优点是不需要事先定义聚类簇的个数，并且收敛性较非层次聚类算法好，能够得到全局最优聚类结果。通过事件检测得到 k 个类，每个事件类都由一组官微突显词博文簇组成，从中选取描述事件类的突显词作为事件的检测结果。

5 实验与分析

实验数据采用第三方数据软件抓取的新浪微博官微数据。目前，还欠缺官微数据研究的标准库，笔者从北京大学、内蒙古科技大学团委、包头交警、南开大学等官微历史数据集中各取 2 000 条，共 8 000 条微博博文作为测试数据并进行人工标注。时间段 t 大小设置为 15 天，对其中的官微博文进行事件的检测。实验环境为 Intel(R) Core(TM) i5-3470 CPU @ 3.2GHz、RAM 为 8GB，操作系统为 64 位的 Microsoft Windows 7。

5.1 官微相关语料库

运用软件抓取官微组织的微博记录，并进行分词等预处理，通过机器深度学习 Word2Vec 训练得到官微相关词和官微组织的相关度权值。北京大学官微相关词的相关度权值如表 1 所示。

表 1 北京大学官微相关度权值

相关词	相关度权值	相关词	相关度权值
北大	0.511464	北京大学第三医院	0.418483
许智宏	0.483764	荣获	0.418440
清华大学	0.470910	生命科学	0.416257
招生办	0.470327	大讲堂	0.416236
深圳	0.468520	展开	0.414171
携手	0.466221	院长	0.411467
揭晓	0.461243	来访	0.409098
代表团	0.451333	团委	0.408964
天文	0.450333	北京大学法学院	0.405279
电视台	0.447696	研究院	0.404885
第一届	0.442393	泰王国	0.404662
代表队	0.440339	物理	0.404421
孔庆东	0.433270	邓宏魁	0.400969
研究生会	0.431442	空间科学	0.398369
6 月	0.421351	博雅	0.398264
研究生院	0.420898	学生会	0.397033

如果官微博文中含有招生办、孔庆东、生命科学、博雅等词时，这些词的 $REL_{i,j}$ 会被赋予较大权重，在突显词检测时会被检测为突显词，作为描述官微事件的突显词用于检测官微事件。另外，对于官微相关词中出现的“电视台”、“6 月”等相关度权值排名问题，将初次得到的相关度权值与官微简介中或标签中的主题词进行二次比对，再次计算其相似度值消除误差。

5.2　突显词提取结果与分析

实验语料采用面向官微历史记录数据集中 200 条博文(每个官微选取 50 条), 人工标注出每条博文中所含的官微突显词, 每条最多标注 10 个官微突显词, 如例中微博 P2 中的突显词: “朗润园”、“北京大学”、“智库”、“林建华”、“校长”等词, 微博 P1 中不含官微突显词, 可以直接被过滤掉。对这 200 条博文, 利用本文所提出的方法, 每条博文提取出 10 个突显性最高的词。通过计算候选突显词在标注词上的准确率、召回率和 F 值评测方法的有效性。实验中分别用 TF-IDF、TextRank 为每条博文提取出 10 个候选词与本文的方法作对比。

突显词实验对比结果如图 1 所示。可以得出, 本文方法相对于改进的 TF-IDF 和 TextRank 方法在准确率、召回率和 F 值方面有显著提高, 在官微数据中提取突显词结果上占有一定优势。对比方法中 TF-IDF 评测结果表明单一考虑词频在官微数据中获取突显词的不足, 评测值很低。TextRank 方法是基于图模型的算法, 相对于 TF-IDF 方法在实验结果中各项值均有所提高, 然而对比本文方法其评测值仍然不足, 表明博文转发、评论等特有属性缺失对实验结果有影响。两种对比方法结果的评测值都较低, 同时也说明官微相关语料、博文影响力等特征权重的加入对特征词突显性检测所起的作用, 使官微相关词的权重明显得到提高, 官微特征词的突显性更加突出, 为下一阶段描述官微事件检测提供了基础支撑。

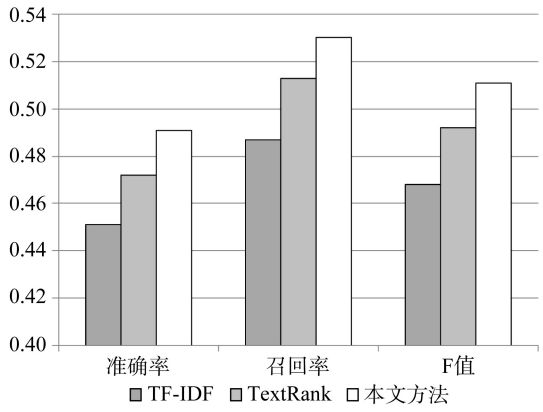

图 1　突显词评价结果

5.3　事件检测结果与分析

由于本文实验环境下, 无法获取某一时间段内现实生活中官微事件总数全部, 无法直接得到传统的召回率。因此由实验室的三位同学对官微事件进行人工标注, 将人工标注的官微事件总数等同为现实中官微事件总数, 将识别到的事件数和人工标注的事件数交集作为识别正确的事件数, 进而计算官微事件的准确率、召回率、F 值。事件检测评价方法如公式(10)–公式(12)所示。

$$准确率=\frac{识别正确的事件数}{识别到的事件数} \tag{10}$$

$$召回率=\frac{识别正确的事件数}{人工标注的事件总数}\propto\frac{识别正确的事件数}{现实中正确事件总数} \tag{11}$$

$$F值=\frac{2\times准确率\times召回率}{准确率+召回率} \tag{12}$$

官微突显词能够表征官微事件相关博文, 而官微事件相关博文又能够反映官微事件, 因此官微突显词检测的正确性直接影响官微事件检测的准确率。实验中设置突显词检测阈值 σ 为 1.85, 得到的突显词效果最好。相似度参数 α_1、α_2、α_3、α_4 参数值分别设置为 0.3、0.3、0.25、0.15 时, 采用凝聚层次聚类算法对 2015 年 10 月 7 日–2016 年 4 月 20 日之间的北京大学官微博文数据进行事件检测, 得到三个具有代表性的官微事件类簇, 以每个类簇中对应权值较大的若干突显词描述该事件类, 如表 2 所示。

将文献[3]中提出的基于突发词聚类的微博事件检测算法作为文献方法, 将文献[5]中提出的方法作为基础方法。将基础方法、文献方法和本文方法在相同的实验环境下进行准确率、召回率与 F 值对比, 实验数据为官微博文数据。

文献方法的实验中, 突显性权重阈值 ϱ 取 2.0, 聚类阈值采用的是距离阈值, 其中增量聚类距离阈值 μ 分别取 300-900 进行实验, 当距离阈值取 500 时, 聚类结果 F 值达到最优值 0.701, 准确率为 0.597, 召回率为 0.850。与本文方法的准确率 0.635、召回率 0.855 及 F 值 0.730 比对, 如图 2 所示。

从图 2 可以看出, 与基础方法和文献方法相比, 本文方法在事件检测结果中的准确率、召回率以及 F 值均有所提高。其中准确率和 F 值提高较明显。分析表明官微突显词检测对官微事件检测准确率有一定的影响。表 2 数据也可以说明官微博文中的突显词能够较清楚地描述官微事件, 起到检测官微事件的作用。同时, 凝聚层次聚类算法收敛性较好, 对于突显词博文聚类的实验效果也具有可观性。

表 2　官微事件检测结果

官微事件	事件类描述	突显词博文聚类	日期
屠呦呦获诺奖，北大师生表示祝贺	屠呦呦 校友诺奖 林建华 校长 北大 医学部 席谈	【林建华校长看望诺贝尔奖获得者屠呦呦校友】10 月 6 日下午，2015 年诺贝尔生理学或医学奖获得者、北京大学校友屠呦呦的家里暖意融融。北大校长林建华一行向屠呦呦校友表示祝贺...	2015-10-7 11: 44: 35
		【踏实做事 献身科学——屠呦呦校友获诺奖后医学部师生一席谈】在校友屠呦呦获得诺贝尔奖后，北大医学部...	2015-10-17 13: 12: 17
空间科学院教授获国家技术发明奖	北大 国家 技术奖 2015 空间科学院 晏磊	#北大新闻#【简讯：北京大学 13 项成果喜获 2015 年度国家科学技术奖】1 月 8 日上午。人民大会堂举行 2015 年度国家科学技术奖励大会。北京大学...	2016-1-8　18: 28: 49
		#科研动态#【地球与空间科学学院晏磊教授获国家技术发明奖二等奖】1 月 8 日，中共中央、国务院在人民大会堂举行 2015 年度国家科学技术奖励大会...	2016-1-16 10: 30: 03
总理来访北大	总理 北京大学 朗润园 智库 林建华 校长 母校 光华管理 农园食堂	#总理来啦# 第一站，克强总理来到位于朗润园的国家发展研究院，了解北京大学智库建设以及国家发展研究院的发展情况。北京大学校长林建华...	2016-4-15 15: 48: 00
		#总理来啦# 第三站，克强总理来到本科期间(1978-1982 年)就读的法学院...光华管理学院的同学们热烈欢迎总理回到母校，总理与同学们合影留念。	2016-4-15 16: 30: 29
		#总理来啦# 夜幕渐渐降临，克强总理一行来到北京大学农园食堂...克强总理在同学们的簇拥下走出农园食堂...	2016-4-15 20: 09: 01

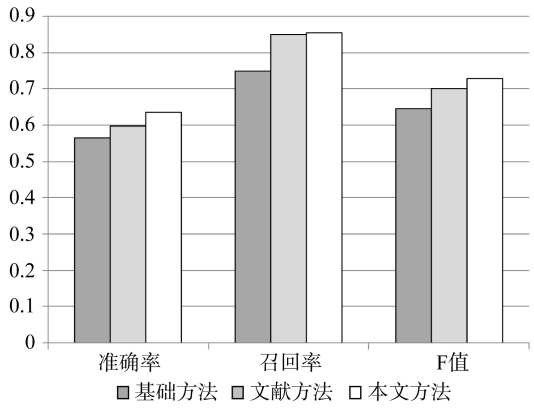

图 2　实验结果对比

基础方法和文献方法准确率偏低可能是由于两种方法获取的突显词不能够很好地表征官微事件，事件噪声较大。在召回率方面，基础方法参数值较低，文献方法和本文方法差距不大，可能是新闻话题抽取方法对于官微数据的敏感性较一般微博事件提取方法的敏感性高，事件检测准确率和召回率的差别得到了不同的 F 值。综上，基础方法和文献方法对于特征词的官微相关性考虑都欠缺，说明在官微这个特殊领域，利用官微突显词博文聚类方法对于官微事件检测结果的有效性，能够将描述官微事件的博文突显词准确地提取出来并用于描述官微事件类，完成官微事件检测。在以后的实验中还要进一步检测本文方法在一般微博中的适用性，确定官微数据和一般微博数据差别所带来的影响。

6　结　语

本文提出一种结合博文影响力、词基础权重和官微相关词权重的官微突显词检测方法，可以对官微博文中的突显词进行准确提取，将描述官微事件的突显词权重增大。通过凝聚式层次聚类算法聚类得到事件类，并用突显词簇描述官微事件作为官微事件检测的结果，取得了可观的效果。该方法不足之处在于，用官微博文历史记录训练官微相关词会出现数据冷启动问题，理论分析可知官微历史记录时间跨度越长记录数量越大，机器学习到的官微相关词就越精准，数据冷启动问题的应对可用大量官微组织新闻语料代替官微历史记录进行训练，以实现官微相关词较准确提取和权重赋值。官微数据的研究工作目前还比较欠缺，官微领域仍是一个待发掘的领域，下一步将建立更加完善的官微相关语料测试库，也将根据官微的特征提出一系列适合官微数据处理的方法，更好地挖掘这个领域存在的价值。

参考文献：

[1] 戴天，吴渝，雷大江．利用组合模型生成微博热点话题事件摘要[J]．计算机应用研究，2016，33(7)：2026-2029．(Dai Tian, Wu Yu, Lei Dajiang. Hot Topic Summarization on Microblog Generated by Model Combination[J]. Application Research of Computers, 2016, 33(7): 2026-2029.)

[2] 贺敏，杜攀，张瑾，等．基于动量模型的微博突发话题检测方法[J]．计算机研究与发展，2015，52(5)：1022-1028．(He Min, Du Pan, Zhang Jin, et al. Microblog Bursty Topic Detection Method Based on Momentum Model[J]. Journal of Computer Research and Development, 2015, 52(5): 1022-1028.)

[3] 郭跃秀，吕学强，李卓．基于突发词聚类的微博突发事件检测方法[J]．计算机应用，2014，34(2)：486-490．(Guo Yixiu, Lyu Xueqiang, Li Zhuo. Bursty Topics Detection Approach on Chinese Microblog Based on Burst Words Clustering[J]. Journal of Computer Applications, 2014, 34(2): 486-490.)

[4] 童薇，陈威，孟小峰．EDM：高效的微博事件检测算法[J]．计算机科学与探索，2012，6(12)：1076-1086．(Tong Wei, Chen Wei, Meng Xiaofeng. EDM: An Efficient Algorithm for Event Detection in Microblogs[J]. Journal of Frontiers of Computer Science and Technology, 2012, 6(12): 1076-1086.)

[5] 郑斐然，苗夺谦，张志飞，等．一种中文微博新闻话题检测的方法[J]．计算机科学，2012，39(1)：138-141．(Zheng Feiran, Miao Duoqian, Zhang Zhifei, et al. News Topic Detection Approach on Chinese Microblog[J]. Computer Science, 2012, 39(1): 138-141.)

[6] Gorling R. A Preliminary Study of Tweet Summarization Using Information Extraction[C]// Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013.

[7] Chakrabarti D, Punera K. Event Summarization Using Tweets[C]//Proceedings of the 15th International AAAI Conference on Weblogs and Social Media.2011.

[8] Li C, Sun A, Datta A. Twevent: Segment-based Event Detection from Tweets[C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York: ACM, 2012: 155-164.

[9] 杨文漪．面向微博的事件检测算法研究[D]．北京：北京邮电大学，2013．(Yang Wenyi. Research on Event Detection Algorithm for Microblog[D]. Beijing: Beijing University of Posts and Telecommunications, 2013.)

[10] 宁瑞芳，欧阳宁，莫建文．基于光流法的聚众事件检测[J]．计算机工程与应用，2012，48(3)：198-201．(Ning Ruifang, Ouyang Ning, Mo Jianwen. Detection of Gathering Events Based on Optical Flow [J]. Computer Engineering and Applications, 2012, 48(3): 198-201.)

[11] 唐明，朱磊，邹显春．基于 Word2Vec 的一种文档向量表示[J]．计算机科学，2016，43(6)：214-217．(Tang Ming, Zhu Lei, Zou Xianchun. Document Vector Representation Based on Word2Vec[J]. Computer Science, 2016, 43(6): 214-217.)

[12] Becker H, Naaman M, Gravano L, et al. Selecting Quality Twitter Content for Events [C]//Proceedings of the 15th International AAAI Conference on Weblogs and Social Media.2011.

[13] Duan Y, Chen Z, Wei F, et al. Twitter Topic Summarization by Ranking Tweets Using Social Influence and Content Quality[C]//Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012). 2012: 763-780.

[14] Mihalcea R, Tarau P. TextRank: Bringing Order into Texts[C]// Proceedings of the 2004 Conference on Empirical Methods in Natural Language.2004: 404-411.

[15] 余珊珊，苏锦钿，李鹏飞．基于改进的 TextRank 的自动摘要提取方法[J]．计算机科学，2016，43(6)：240-247．(Yu Shanshan, Su Jindian, Li Pengfei. Improved TextRank-based Method for Automatic Summarization[J]. Computer Science, 2016, 43(6): 240-247.)

[16] 朱征宇，孙俊华．改进的基于知网的词汇语义相似度计算[J]．计算机应用，2013，33(8)：2276-2279．(Zhu Zhengyu, Sun Junhua. Improved Vocabulary Semantic Similarity Calculation Based on HowNet [J]. Journal of Computer Applications, 2013, 33(8): 2276-2279.)

[17] 黄承慧，印鉴，侯昉．一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]．计算机学报，2011，34(5)：856-864．(Huang Chenghui, Yin Jian, Hou Fang. A Text Similarity Measurement Combining Word Semantic Information with TF-IDF Method[J]. Chinese Journal of Computers, 2011, 34(5): 856-864.)

作者贡献声明：

高永兵，杨贵朋：提出研究思路，设计研究方案，论文起草及最终版本修订；

高永兵，杨贵朋，张娣，马占飞：进行实验和测评；

马占飞：实验分析。

利益冲突声明：

所有作者声明不存在利益冲突关系。

研究论文

支撑数据:

支撑数据由作者自存储, E-mail: gaoyongbing@126.com。

[1] 高永兵, 杨贵朋. weibo_corpus.bin. 官微相关语料集.

[2] 高永兵, 杨贵朋. data_weibo.xlsx. 官微博文数据记录表.

[3] 高永兵, 杨贵朋. data_result.xlsx. 数据处理结果表.

收稿日期: 2017-04-05
收修改稿日期: 2017-05-27

Detecting Events from Official Weibo Profiles Based on Post Clustering with Burst Words

Gao Yongbing[1] Yang Guipeng[1] Zhang Di[1] Ma Zhanfei[2]

[1](School of Information Engineering, Inner Mongolia University of Science and Technology, Baotou 014010, China)

[2](Department of Computer, Baotou Teachers' College, Baotou 014010, China)

Abstract: [**Objective**] This paper aims to remove the unrelated information from the official Weibo (micro-blog) profiles, and then retrieves the posts on official events. [**Methods**] First, we used the word2vec machine learning model to train the official Weibo datasets. Then, we proposed an official micro burst words detection method based on the influence of Weibo posts, the base weight and the related official profiles. Third, we calculated the similarity of blog posts with the burst words, and used hierarchical clustering algorithm to select burst words for the target events. [**Results**] The proposed algorithm had better precision (63.5%), recall (85.5%) and F values (0.73) than the traditional TF-IDF and TextRank algorithms. [**Limitations**] The official profiles did not have enough historical data on the events. [**Conclusions**] The burst words help us detect official events effectively from the official Weibo profiles.

Keywords: Official Micro-blog Related Words Burst Words Official Microblog Events Word2Vec